Analysis of binaural cue matching using ambisonics to binaural decoding techniques

Analysis of Binaural Cue Matching using Ambisonics to Binaural Decoding Techniques

B. Wiggins[1]

[1] *University of Derby, UK, Email: b.j.wiggins@derby.ac.uk*

Abstract

Last year Google enabled spatial audio in head-tracked 360 videos using Ambisonics to binaural decoding on Android mobile devices. There was some early criticism of the 1st order to binaural conversion employed by Google, in terms of the quality of localisation and noticeable frequency response colouration. In this paper, the algorithm used by Google is discussed and the Ambisonics to Binaural conversion using virtual speakers analysed with respect to the resulting inter-aural time, level, and spectrum differences compared to an example HRTF data set. 1st to 35th order Ambisonics using multiple virtual speaker arrays are implemented and analysed with inverse filtering techniques for smoothing the frequency spectrum also discussed demonstrating 8th order decoding correctly reproducing binaural cues up to 4 kHz.

Introduction

In April 2016, Google enabled head/phone tracked Spatial Audio to be uploaded and auditioned on the YouTube platform (over one year after 360, head/phone tracked, videos were enabled on the platform). Google chose Ambisonics (using the ambiX standard [1]) for the format used to encode the 360 degree sound scene. Initially, excitement and a surge of interest in Ambisonics occurred, but soon after, on-line forums and discussion boards voiced issues regarding localisation quality and frequency response, compared with the original recording. In this paper, a discussion of the implementation and limitations of the format will be discussed focussing on the differences between inter-aural differences present between measured, and Ambisonically synthesised HRTFs.

Ambisonics to Binaural

Converting Ambisonics B-Format to binaural audio for headphones is well documented with McKeag and McGrath using 1st order Ambisonic recordings to feed head tracked binaural audio over headphones in 1996 [2]. The auralisation of any loudspeaker based system can be achieved by convolving the HRIRs of the location of the loudspeakers with the audio fed to those loudspeakers which will result in 2 x N convolutions (where N is the number of loudspeakers). The polar patterns of the signals for a Furse-Malham channel ordering and normalised B-Format signals are shown below in Figure 1.

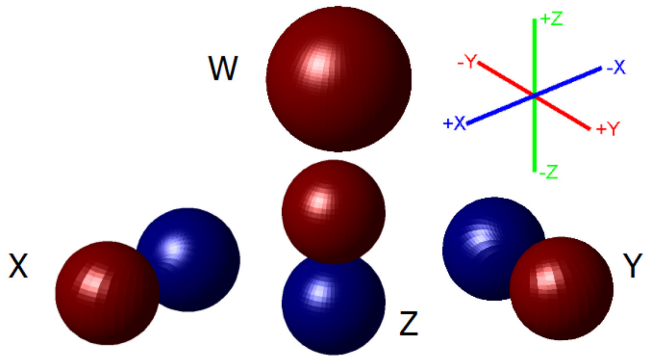

Figure 1: 1st Order B-Format Signal Polar Patterns (Furse-Malham Scheme)

The signals contained within these channels, representing the full 360 degree sound scene, can be transformed into signals fed to loudspeakers by deriving an Ambisonic decoder and binauralising. The resulting binaural output will be (at its simplest) a linear combination of these 4 channels (for 3D) or 3 channels (omitting the Z for 2D) to create the speaker feeds, followed by the convolution of each speaker feed with the corresponding HRIR. The left and right ear results can then be summed to the left and right channels of the headphones as shown in the block diagram in Figure 2.

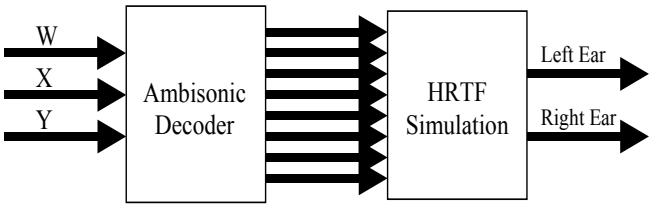

Figure 2: 2D B-Format to Binaural Conversion Process **[3]**

One benefit of Ambisonics is that the processes involved in both the Ambisonic Decoder *and* HRTF simulation parts can be rolled up into one operation resulting in a pair of HRIRs for each B-Format channel [3]. This reduces the number of convolutions needed to one per B-Format channel, no matter how many speakers are to be auralised.

Mathematically, once an Ambisonic decoder has been designed (which will be a matrix of coefficients generated by inverting the spherical harmonic weightings of the loudspeaker directions), the 1st order HRIRs can be calculated as shown in equation (**1**). This example assumes 8 speaker locations with corresponding HRIRs

$$W^{hrir} = \sum_{c=1}^{8} Wcoef_c \times HRIR_c$$
$$X^{hrir} = \sum_{c=1}^{8} Xcoef_c \times HRIR_c \quad (1)$$
$$Y^{hrir} = \sum_{c=1}^{8} Ycoef_c \times HRIR_c$$

Where c represents loudspeaker number

Once the HRIRs have been generated, the binaural output can be derived as shown using the processing in equation (**2**). If the HRIRs are left/right symmetrical (that is, the HRIRs for a source at 30 degrees, is the same as the HRIRs for -30 degrees, but with the left and right responses swapped), then this can be further reduced to a single convolution per spherical harmonic [3].

$$\begin{aligned} Left &= \left(W \otimes W_L^{hrir}\right) + \left(X \otimes X_L^{hrir}\right) \\ &\quad + \left(Y \otimes Y_L^{hrir}\right) \\ Right &= \left(W \otimes W_R^{hrir}\right) + \left(X \otimes X_R^{hrir}\right) \\ &\quad + \left(Y \otimes Y_R^{hrir}\right) \end{aligned} \tag{2}$$

$$W, X \text{ and } Y \text{ are } B-Format\ Audio$$

$$W, X, Y^{hrir} \text{are the calculated hrirs}$$

$$_L \text{and}_R \text{denote } Left \text{ and } Right \text{ ear}$$

For more detail of the current state of the art regarding spherical harmonics and their use with binaural audio, the reader is directed to the concise summary given by Politis and Poirier-Quinot [4].

YouTube Spatial Audio Frequency Correction

Soon after YouTube enabled spatial audio, the author measured the 1st order Ambisonics to Binaural filters by uploading a video that contained a swept sine wave [5] on each of the B-Format channels sequentially and recording the binaural result from a supported android phone. The result was then convolved with the inverse filter to obtain the left and right HRIRs for the W, X, Y and Z channels. The filters obtained can be seen (for a phone orientation of 0 and 90 degrees) in Figure 3 and Figure 4, respectively.

The fact that the IRs for three of the four channels are identical, and one is identical, but inverted, for both the left and right responses show that the YouTube filters are using HRIRs that assume left/right symmetry. The fact that the impulses are identical with the phone oriented at 0 and 90 degrees also suggests that YouTube is rounding the rotation angle as it's very unlikely that the manual phone rotation hit exactly 90 degrees! The frequency response of these filters is also shown in Figure 5. Note YouTube filters audio above around 16.4kHz.

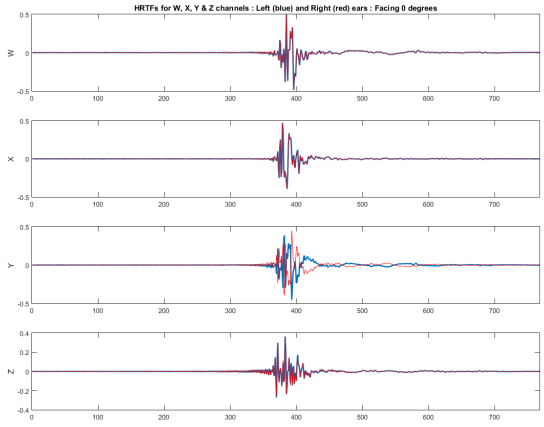

Figure 3: Measured HRIRs from YouTube with the phone at 0 degrees orientation

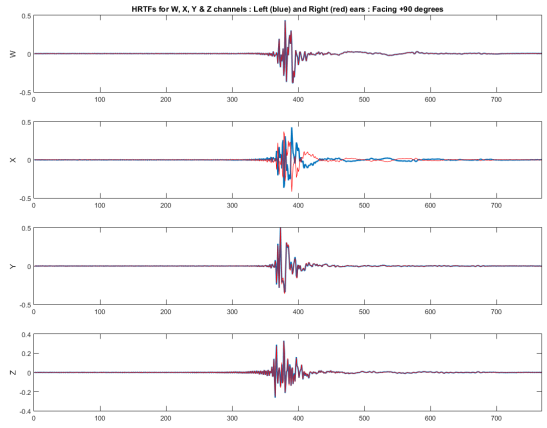

Figure 4: Measured HRIRs from YouTube with the phone at +90 degrees orientation

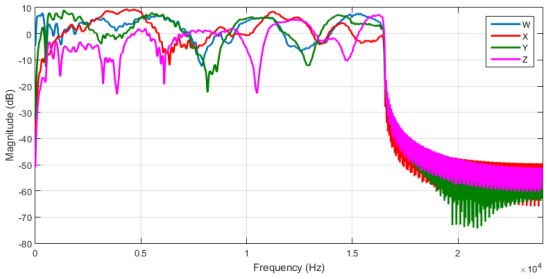

Figure 5: Frequency response of measured HRIRs from YouTube.

One initial comment that started to emerge in discussion boards and forums was that the YouTube implementation sounded quite coloured, in terms of its frequency response. For example, the resulting responses of an Ambisonic source panned at 90 degrees to the left of a listener can be seen below in Figure 6.

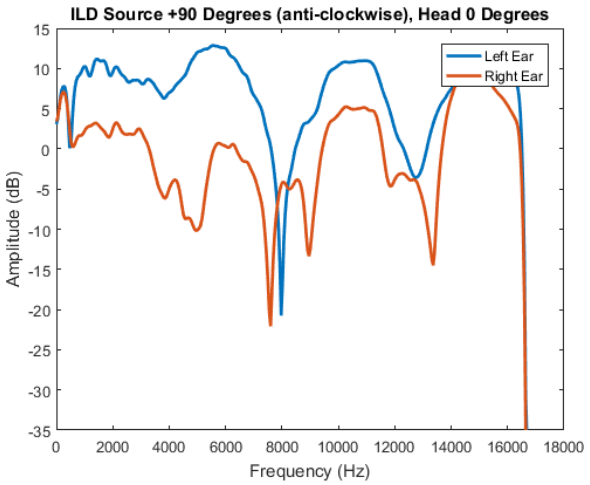

Figure 6: Frequency response of measured HRIRs from YouTube.

The filtering present is due to the response of the head and torso used in the measurement of the HRTFs and possibly the speaker and microphones used as well. If the head and torso matched exactly your own, then this filtering would not be noticed and, instead, interpreted as directional information, but as it doesn't, it will be perceived as colouration in the

frequency response. In order to equalise the system, a correction EQ curve can be applied to all the Ambisonic channels equally before uploading to YouTube. First, a method do decide on the 'average' response of the system is needed. There could be a few methods for this, but the simple approach used here was to pan a repeated impulse around the listener, storing the resulting IR each time. These responses can then be summed together, and the frequency response averaged (an RMS type approach as shown in equation (3) has been used for in this example). This is then an 'average' response of the system. The system is then inverted (adding delay as it is non-minimum phase) and then the filter is decomposed into its minimum phase only response for the EQ (as that's all we're really interested in and reduces potentially damaging pre-ringing in the filter). The average response and that of its inverse are shown below in Figure 7, with the impulse response of the inverse filter (minimum phase) shown in Figure 8. If this filter is applied to all B-Format channels equally, no corruption of the spatial properties of the recording will occur. A video demonstrating the more natural frequency response reproduction on YouTube before and after application to the B-Format channels can be found at [7]. Anecdotal evidence suggests that this corrected implementation is preferred, and sounds more natural, than the uncorrected version present on YouTube until they, more recently, updated the HRTFs used to the set from the SADIE project [12].

$$Freq(\omega,\theta) = \mathcal{F}\left(W_L^{hrir} + \cos(\theta)\, X_L^{hrir} + \sin(\theta)\, Y_L^{hrir}\right)$$

$$YTResponse(\omega) = \sqrt{\left(\sum_{\theta=0}^{359} |Freq(\omega,\theta)|^2\right)} \quad (3)$$

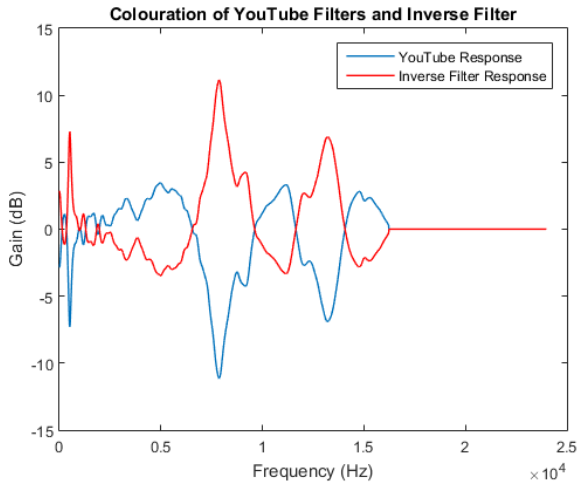

Figure 7: Average frequency response of a source panned horizontally around the listener and its inverse.

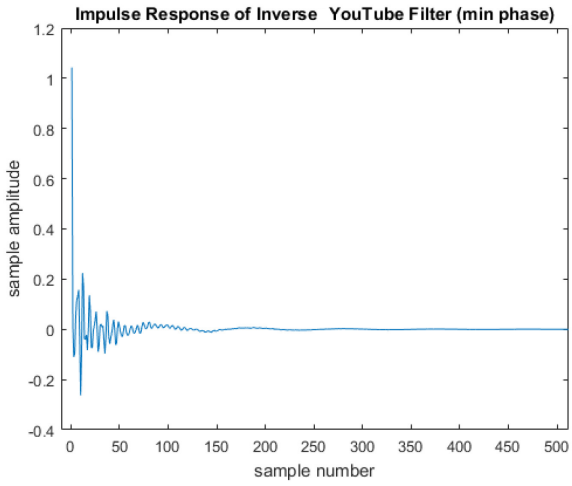

Figure 8: Minimum phase impulse response of the inverse filter shown in Figure 7

Ambisonics to Binarual Inter-aural Cues

Ultimately, the Ambisonics system will be interpreted by the ear/brain system of the listener, meaning it will need to satisfy the auditory cues necessary to convince a listener that the audio is located at the desired direction and distance away. The three more obvious cues to satisfy are inter-aural time difference (ITD – used at low frequencies [< 1kHz]), inter-aural level difference (ILD – most useful at mid frequencies [> 700Hz]) and pinna filtering and monoaural spectral cues (most prevalent at high frequencies). Work from Kearney and Doyle concentrated on spectral/pinna cues in the Ambisonics to binaural decoding with respect to height perception [8]. This work will mainly look at the inter-aural differences in the horizontal plane.

Both the ILD and ITD will be frequency dependent due to the complex structure of the head and torso. The ILD of a set of Head Related Transfer Functions (HRTFs) can be found, with respect to frequency (ω) and angle of incidence (θ) as shown in equation **(4).**

$$\boldsymbol{ILD(\theta,\omega) = 20\log_{10}\left(\frac{|HRTF_L(\theta,\omega)|}{|HRTF_R(\theta,\omega)|}\right)} \quad (4)$$

Similarly, the ITD can be calculated, with respect to frequency (ω) and angle of incidence (θ) using the group delay (rate of change of phase) as shown in equation (5).

$$\boldsymbol{\tau_{d_L}(\omega) = -\frac{d\phi_L(\omega)}{d\omega}\ where\ \phi_L(\omega) = \angle HRTF_L(\theta,\omega)}$$
$$\boldsymbol{\tau_{d_R}(\omega) = -\frac{d\phi_R(\omega)}{d\omega}\ where\ \phi_R(\omega) = \angle HRTF_R(\theta,\omega)} \quad (5)$$
$$\boldsymbol{ITD(\theta,\omega) = \tau_{d_L}(\omega) - \tau_{d_R}(\omega)}$$

The resulting plots for ILD and ITD using the MIT Kemar set of HRTFs are shown below in Figure 9 and Figure 10 respectively.

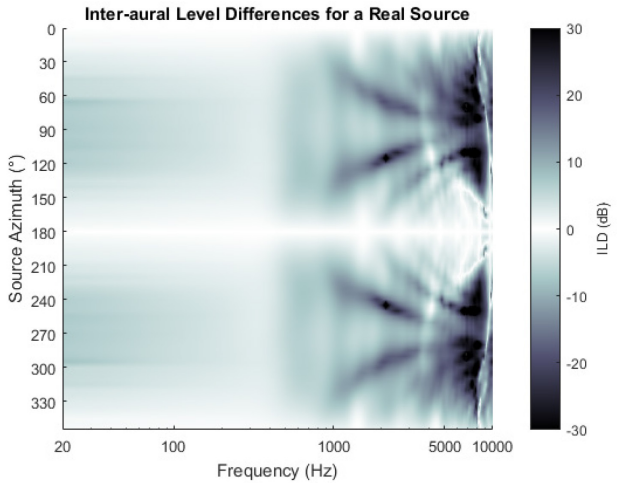

Figure 9: ILD in dB for different source angles and frequencies (up to 10kHz)

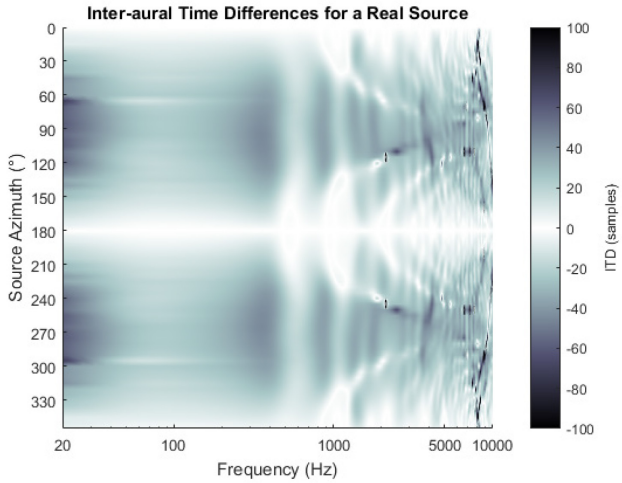

Figure 10: ITD in samples (fs = 44.1kHz) for different source angles and frequencies (up to 10kHz)

The same figure can then be generated for an Ambisonically panned source (for example, Figure 11), and the difference between the real and the Ambisonic cues showing at which frequencies the cues are correct, where they aren't, and how much they deviate from correct (in dB for ILD and samples for ITD).

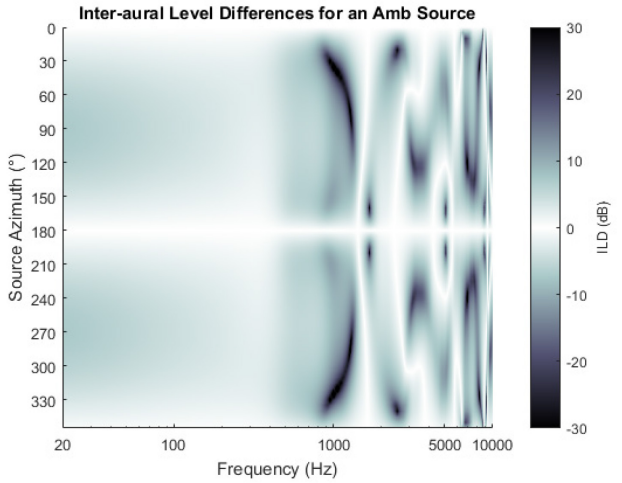

Figure 11: ILD for a 1st Order Ambisonically panned binaural source.

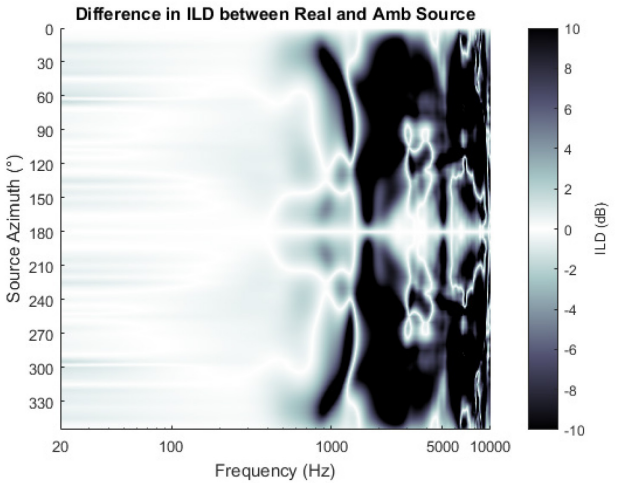

Figure 12: Difference in ILD between a real and a 1st Order Ambisonically panned source (note change in colour scale to amplify errors).

If the differences in ILD in Figure 12 are observed, it can be seen that although there is reduced error at low, compared to high, frequencies, some significant errors are present. These issues can be due to low frequency errors in the measurement of the HRTFs. A new set of low frequency corrected HRTFs have been made available by Erbes, et al. [10] which gives more consistent results and will be used from this point in the analysis (the currently chosen YouTube filters, from the SADIE project [12], also exhibit these issues, incidentally. The perceptual effects of these errors have not yet been investigated). For example, the difference between real and Ambisonic ILD and ITD using the HRIRs from [10] can be seen in Figure 13 and Figure 14 respectively with little low frequency error now present.

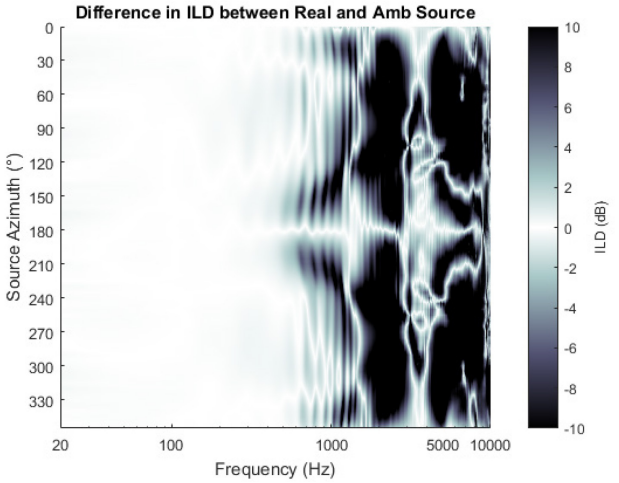

Figure 13: ILD differences between real and Ambisonically panned source using the low frequency corrected HRTF set from **[10].**

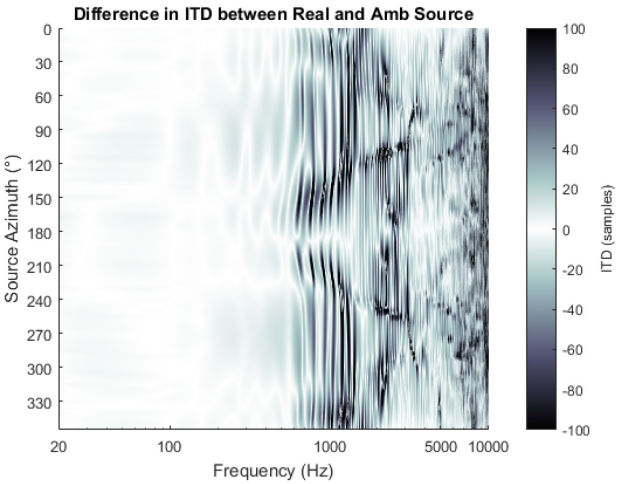

Figure 14: ITD differences between real and Ambisonically panned source using the low frequency corrected HRTF set from **[10].**

These plots are obtained using the virtual loudspeaker array shown in Figure 15. So, what happens if the array is offset so that the four loudspeakers are placed at +/-45° and +/-135° as shown in Figure 16? Figure 17 and Figure 18 show the differences in ILD and ITD for the offset speaker array of Figure 16. From inspection it seems the correctly reproduced area is the same and the only difference is in the response above this frequency. The plot shown in Figure 20, which is the difference between the two difference plots in ITD (in this example) confirms this (the same is found for ILD), thus showing that as long as the HRTFs are correctly measured at low frequencies *and* the number of virtual loudspeakers is enough to correctly sample the system, the differences between the arrays are only observed above the frequency limit of correct operation. Note that this is true even though the calculated B-Format HRIRs are quite different, when viewed in the time domain as shown in Figure 21 and Figure 22.

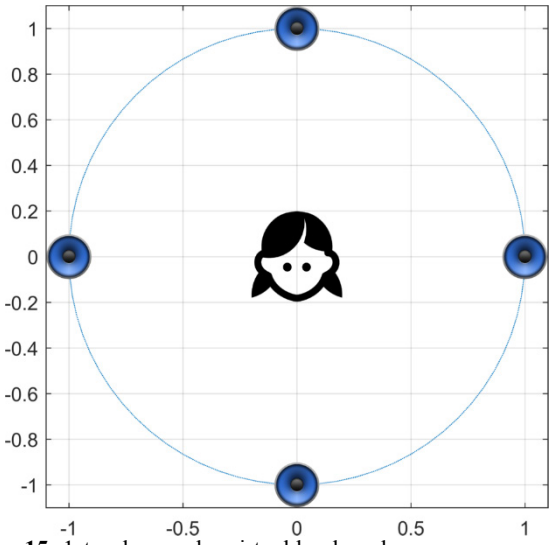

Figure 15: 1st order regular virtual loudspeaker array

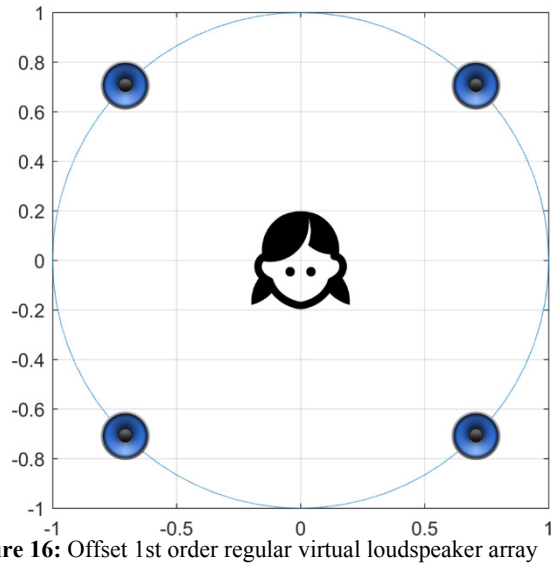

Figure 16: Offset 1st order regular virtual loudspeaker array

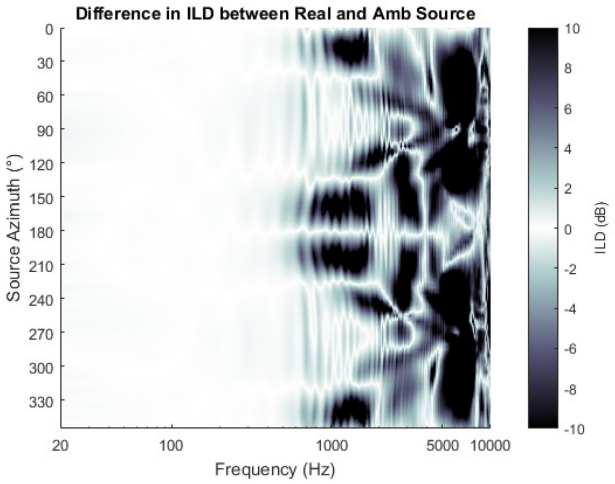

Figure 17: Difference in ILD for an offset virtual speaker array

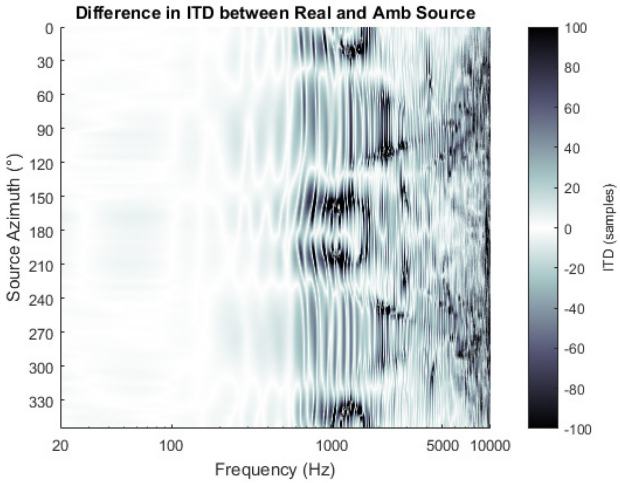

Figure 18: Difference in ITD for an offset virtual speaker array

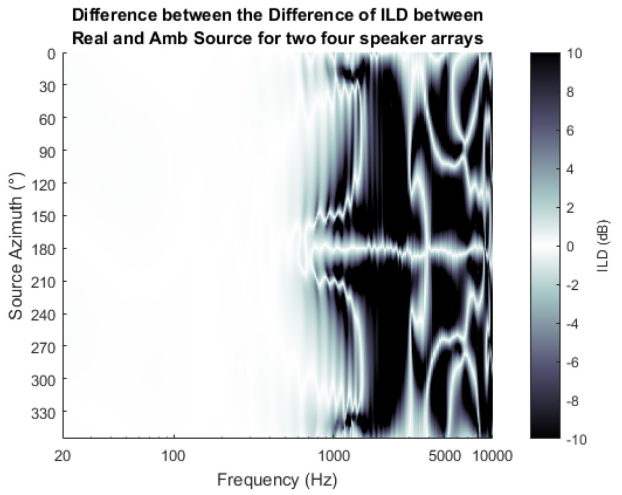

Figure 19: Difference between the ILD difference plots of the two speaker arrays

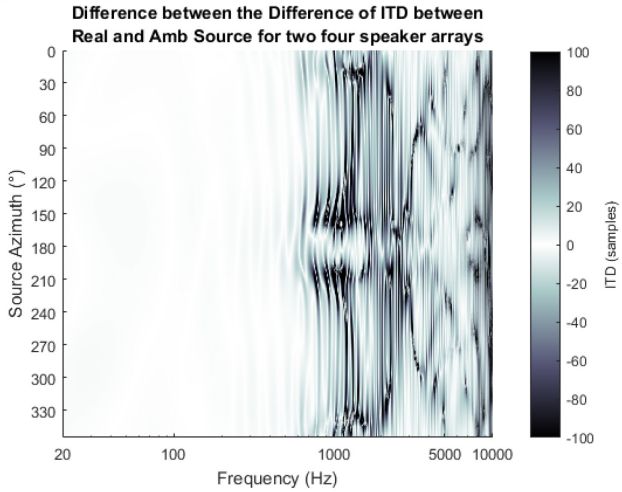

Figure 20: Difference between the ITD difference plots of the two speaker arrays

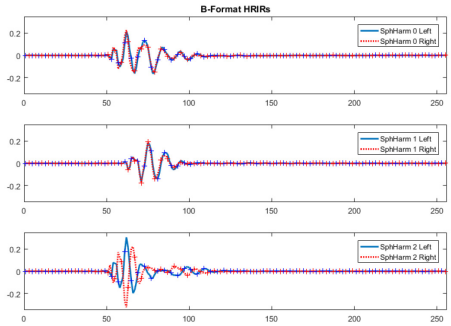

Figure 21: HRIRs for W, X and Y for the speaker array in Figure 15

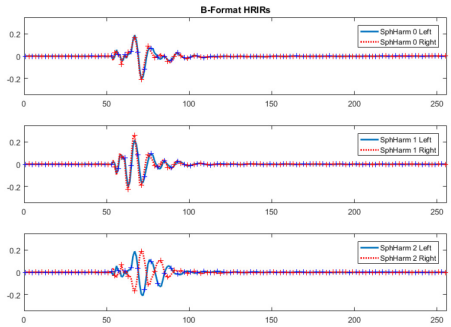

Figure 22: HRIRs for W, X and Y for the speaker array in Figure 16

To improve on this situation, higher order Ambisonics can be implemented (as is already the case in Facebook [2nd order] and Google's Jump Inspector [3rd order]). As demonstrated by Daniel et al [11], as the Ambisonic order is increased, the frequency of correct operation should increase (due to the increased spatial aliasing frequency). The number of speakers (essentially discrete sampling points) needs to increase with the Ambisonic order to correctly reproduce the higher order spherical harmonics. The 2D polar patterns of 1st and higher order components (up to 4th order) can be seen below in Figure 23.

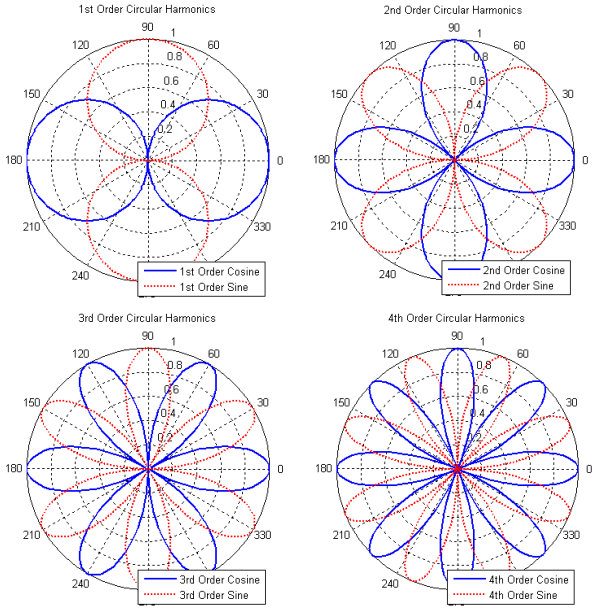

Figure 23: 2D polar patterns of horizontal only higher order B-Format components.

The equations for deriving these harmonics can be seen below in equation (6).

$$S_{mn}(\theta) = \begin{cases} \cos(m\theta), & m > 0 \\ 1, & m = 0 \\ \sin(-m\theta), & m < 0 \end{cases} \quad (6)$$

Where n = ambisonic order
m = +/-n

Increasing the order and the number of speakers (where the number of speakers needed will be 2(n+1) where n is the Ambisonic order), will result in the improvements in ILD and ITD as shown in Figure 24 to Figure 35 from 3rd to 35th order, using 8 to 72 virtual loudspeakers and 7 to 71 B-Format channels respectively. From these plots, the increase in the correctly reproduced frequency limit can clearly be seen with 35th order, 72 loudspeakers, providing correct cues above 10kHz. 35th order requires a measured HRTF pair (or loudspeaker) every 5 degrees, and the limit for many HRTF datasets available. Some insight can also be drawn from the time domain impulse responses of the system. Figure 36 to Figure 40 show the impulse responses of a source at 90 degrees to the listener, both measured and Ambisonically reproduced. 90 degrees was chosen as these will be the IRs with the largest inter-aural differences, both in time and amplitude. Here, we can observe the near-ear response encoded well, with most of the error evident in the far-ear response. It's also worth noting that the error in the

IRs occur when the ear should be experiencing a zero amplitude time delay with the erroneous signal fading as the Ambisonic order increases. This makes sense as the signals are derived as a sum of weighted outputs from all the HRTFs/loudspeakers, even ones at the opposite side of the array. It is these outputs that are minimised as the order increases but, as shown before, will be artefacts above the spatial aliasing frequency.

Conclusions

Many on-line discussions regarding Ambisonics to binaural 3D audio delivery argue between Ambisonics being correct, or not. This paper has demonstrated how Ambisonics to binaural conversion, as utilised in 360 videos and 3D audio for virtual reality is more correctly interpreted as an HRTF pair interpolation method, where the more samples taken (loudspeakers used), and Ambisonic order increased, the higher the frequency that the system provides correct results. 1st order correctly reproducing inter-aural cues up to around 400Hz and 35th order needed to reproduce cues correctly beyond 10kHz. It has also been shown that the locations of the loudspeakers/HRTFs do not matter for the correctly reproduced portion so long as they are regularly spaced and of sufficient number to sample/reproduce the required order of spherical harmonics. Differences will be observed above this frequency, however. While a method for correcting the average frequency response of a Ambisonics to binaural conversion has been presented, no firm conclusions can yet be drawn as only anecdotal evidence has currently been sought. This anecdotal evidence does suggest a preference for the corrected version of the filters, however.

These conclusions lead to several further avenues of investigation. What are the perceptual effects of the errors above the frequency of correct operation, and does changing the decoding scheme (to maximise the energy vector, or use in-phase decoding as discussed in [13]) have any useful perceptual effects for a listener in the sweet spot (as the binaural listener will always be in the centre of the array)? Is it more useful to observe ILD and ITD in frequency bands (as originally carried out in [3])? Would this better match the perception of the system if, say, 3rd octave or octave bands were used? The interaural and monaural cues are only correct to a particular frequency, above this frequency the errors deviate from the measured ITD, ILD and pinna cues. It is suspected that above this frequency, a simple spherical head model would actually provide better, and more natural results in this regard and could be particularly relevant to elevation cues which are often pinna based, affected by higher frequencies than covered by the currently available implementations that use 1st to 3rd order Ambisonics (YouTube uses 1st order, Facebook 2nd order and Google's Jump Inspector, 3rd order with gaming APIs following suit).

References

[1] Christian Nachbar, Franz Zotter, Etienne Deleflie, and Alois Sontacchi, "ambiX - A Suggested Ambisonics Format," in Ambisonics Symposium, Lexington, 2011.

[2] McKeag, A., McGrath, D. (1996) Sound Field Format to Binaural Decoder with Head-Tracking. 6th Austrailian Regional Convention of the AES, Melbourne, Austrailia. 10 – 12 September. Preprint 4302. URL: http://www.aes.org/e-lib/browse.cfm?elib=7477

[3] Wiggins, B. Paterson-Stephens, I., Schillebeeckx, P. (2001) The analysis of multi-channel sound reproduction algorithms using HRTF data. 19th International AES Surround Sound Convention, Germany, p. 111-123. URL: http://www.aes.org/e-lib/browse.cfm?elib=10112

[4] Politis, A. and Poirier-Quinot, D. (2016) JSAmbisonics: A Web Audio library for interactive spatial sound processing on the web. Interactive Audio Systems Symposium, York, UK.

[5] Farina, A. (2000). Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. *Audio Engineering Society*. URL: http://www.aes.org/e-lib/browse.cfm?elib=10211

[6] Wiggins, B. (2016) YouTube 360 VR Ambisonics Teardown. URL: https://www.brucewiggins.co.uk/?p=700

[7] Wiggins, B. (2016). YouTube Spatial Audio Inverse Filter. [online] The Blog of Bruce. URL: https://www.brucewiggins.co.uk/?p=757

[8] Kearney, G & Doyle, T 2015, Height Perception in Ambisonic Based Binaural Decoding. in Audio Engineering Society Convention 139: Papers. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=17979

[9] Gardner, B., Martin, K., (1994) HRTF Measurement of a KEMAR Dummy-Head Microphone, MIT Media Laboratory. URL: http://sound.media.mit.edu/resources/KEMAR.html

[10] Erbes, V., Geier, M., Wierstorf, H. and Spors, S. (2017) Free Database of Low Frequency Corrected Head-Related Transfer Functions and Headphone Compensation Filters. Audio Engineering Society Convention 142. Audio Engineering Society.

[11] Daniel, J., Rault, J., Polack, J. (1998) Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions, preprint no. 4795, 105th Audio Engineering Society Convention

[12] Kearney, G. & Doyle, T., (2015) A Virtual Loudspeaker Database for Ambisonics Research. ICSA 2015: 3rd International Conference on Spatial Audio. Verband Deutscher Tonmeister e.V.,

[13] Daniel, J. (2000). Jérôme Daniel's 3D Sound Research : The Experimenter Corner (Hearing Higher Order Ambisonics). URL: http://gyronymo.free.fr/audio3D/the_experimenter_corner.html

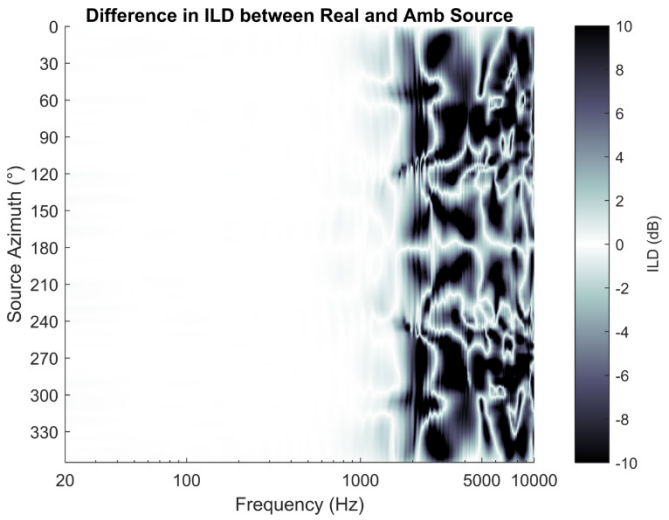

Figure 24: 3rd Order ILD using 8 speakers

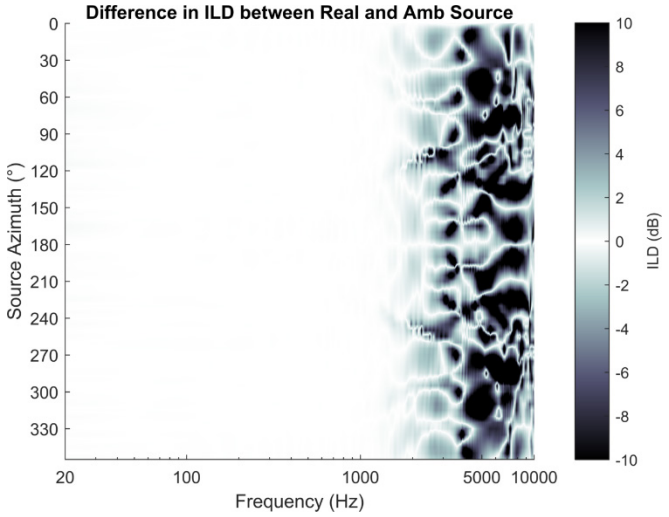

Figure 25: 5th Order ILD using 12 speakers

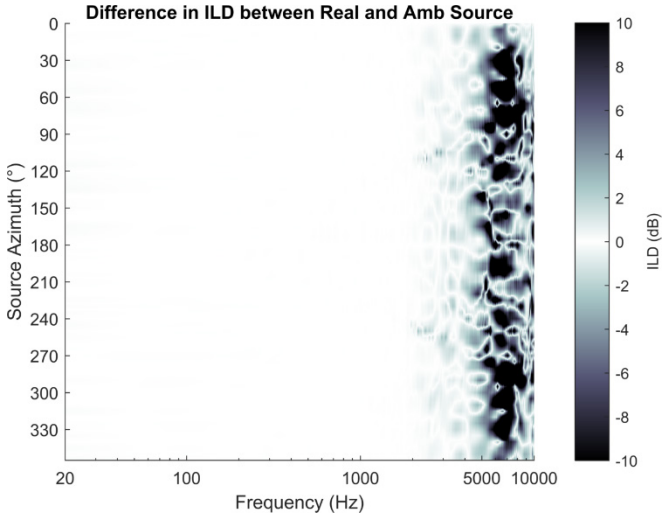

Figure 26: 8th Order ILD using 18 speakers

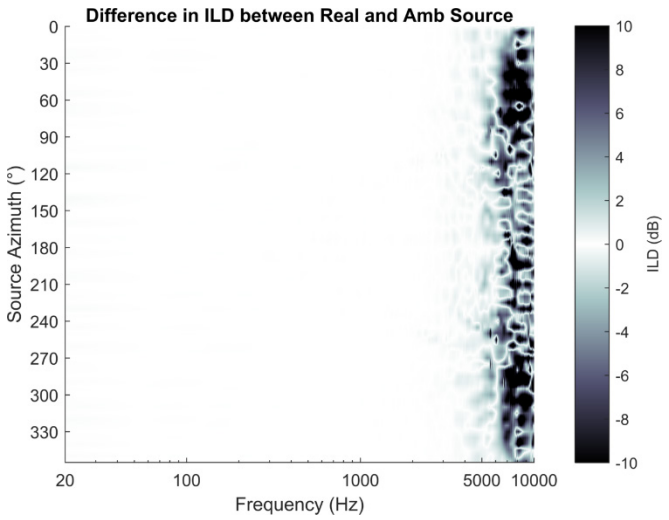

Figure 27: 11th Order ILD using 24 speakers

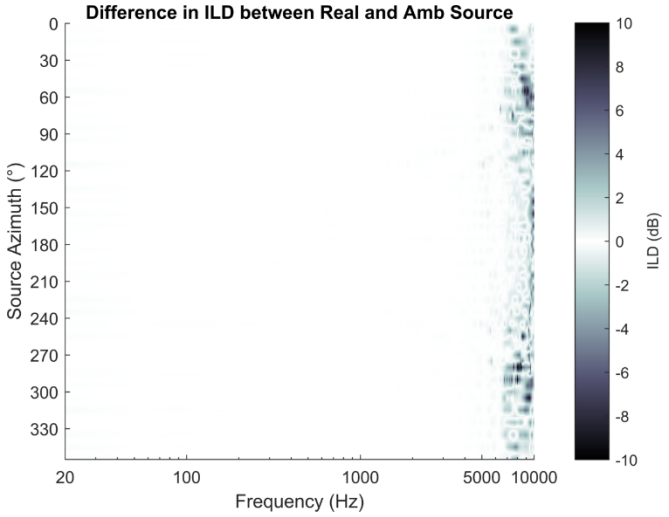

Figure 28: 17th Order ILD using 36 speakers

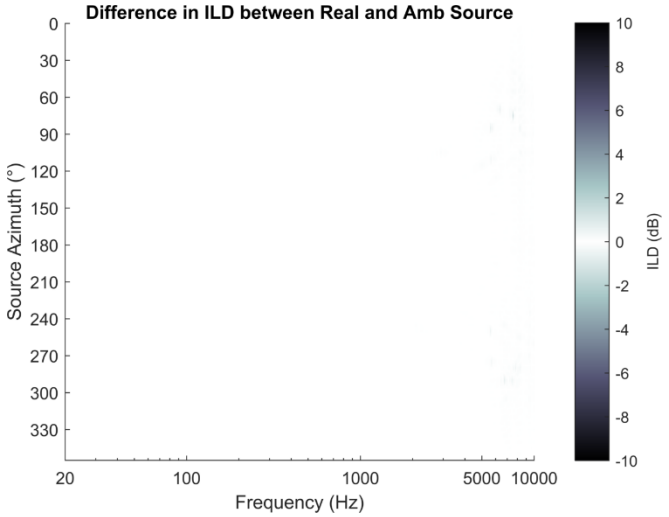

Figure 29: 35th Order ILD using 72 speakers

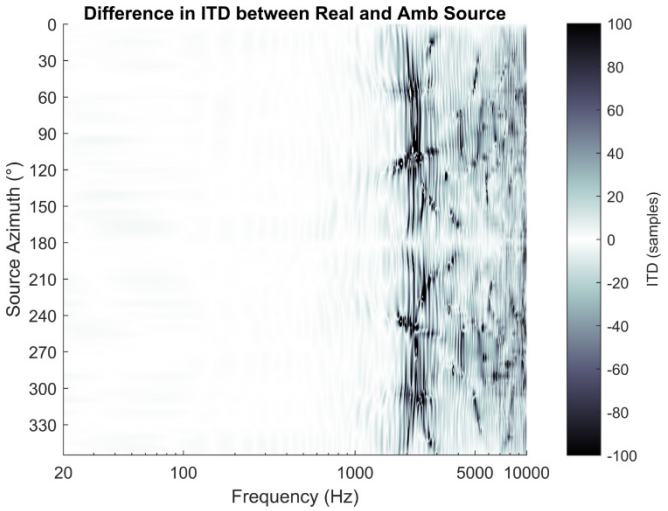

Figure 30: 3rd Order ITD using 8 speakers

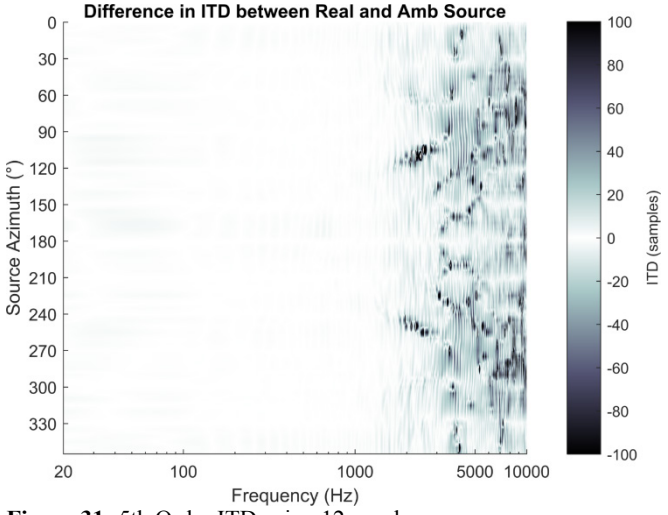

Figure 31: 5th Order ITD using 12 speakers

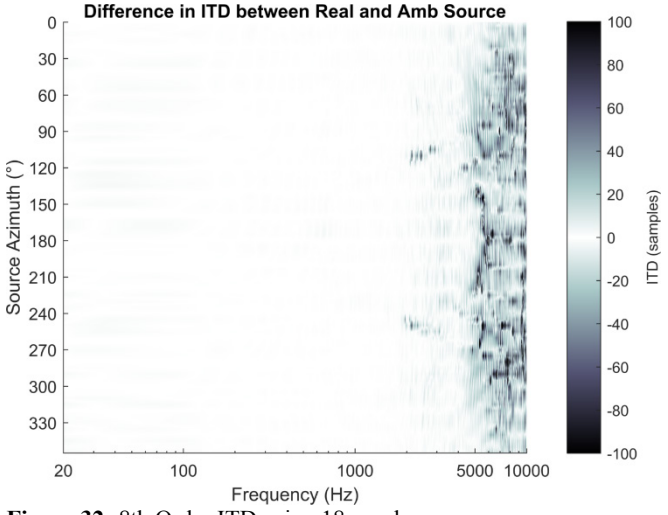

Figure 32: 8th Order ITD using 18 speakers

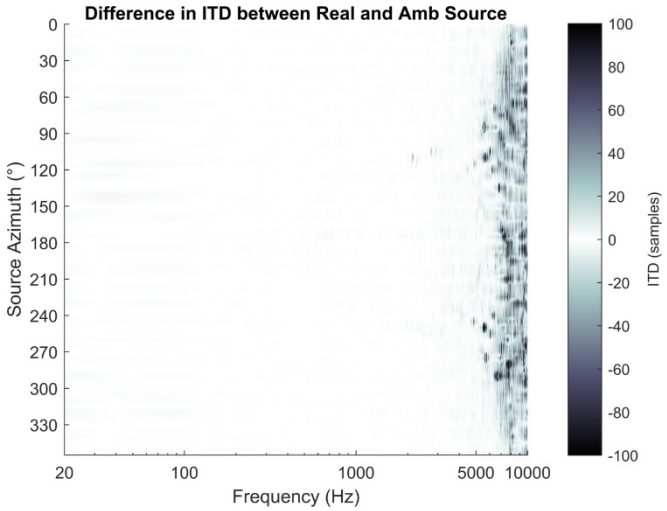

Figure 33: 11th Order ITD using 24 speakers

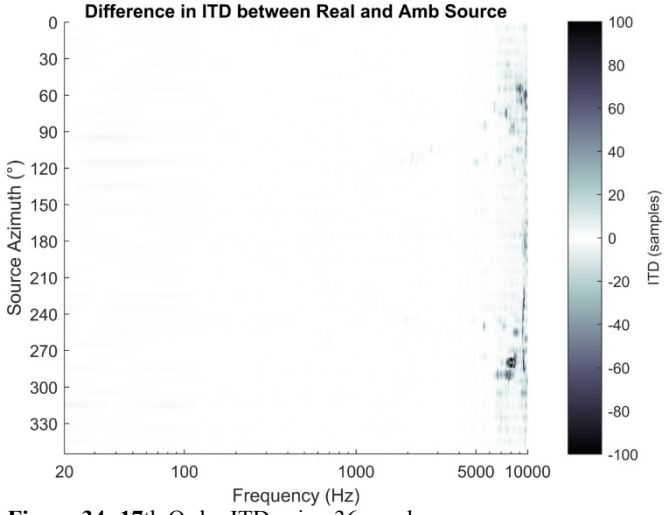

Figure 34: 17th Order ITD using 36 speakers

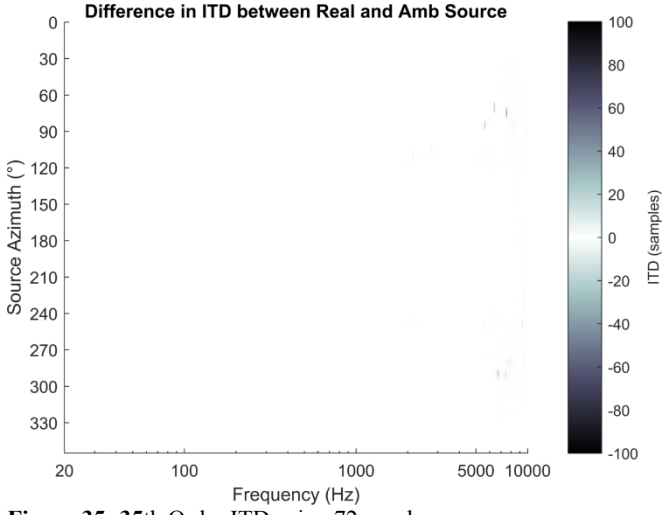

Figure 35: 35th Order ITD using 72 speakers

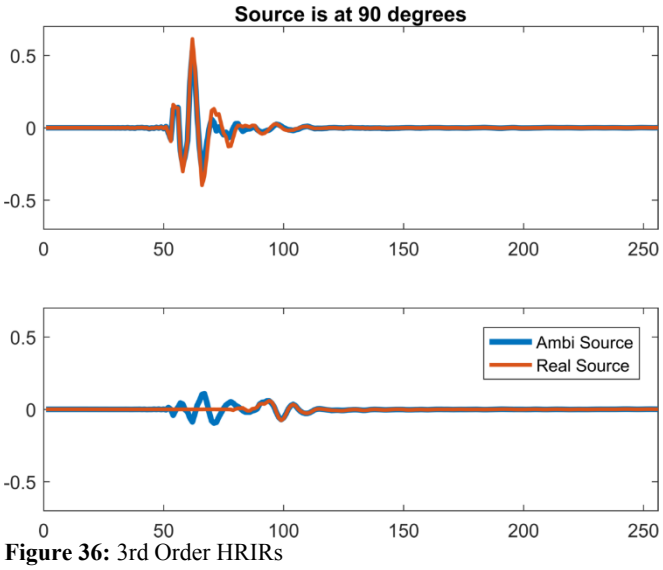

Figure 36: 3rd Order HRIRs

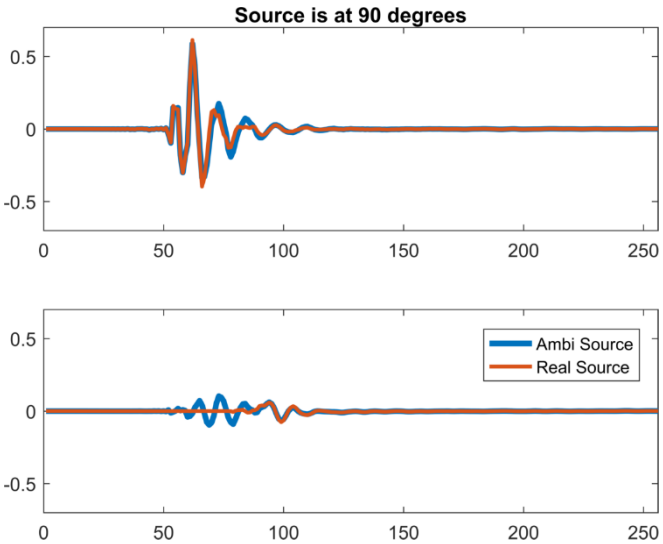

Figure 37: 5th Order HRIRs

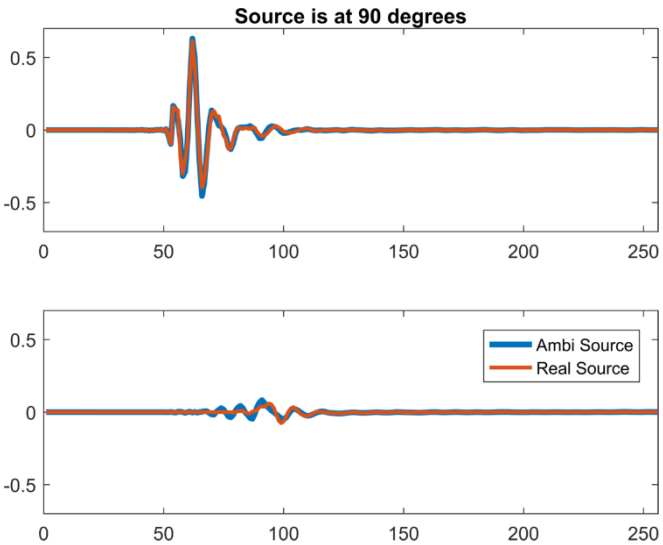

Figure 38: 8th Order HRIRs

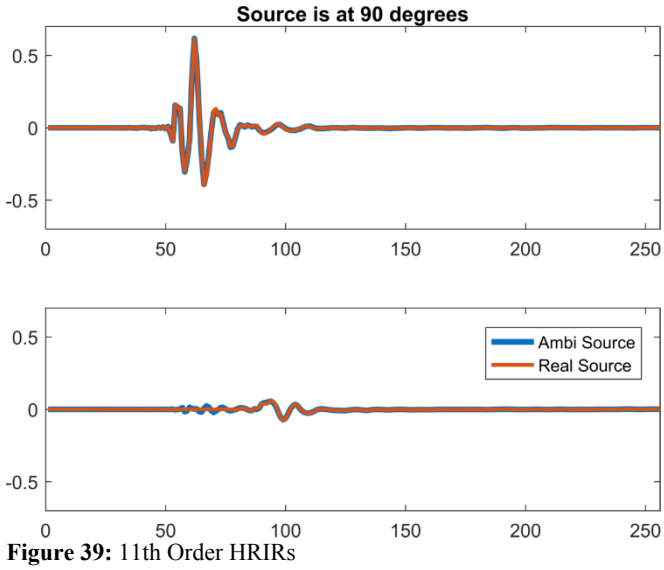

Figure 39: 11th Order HRIRs

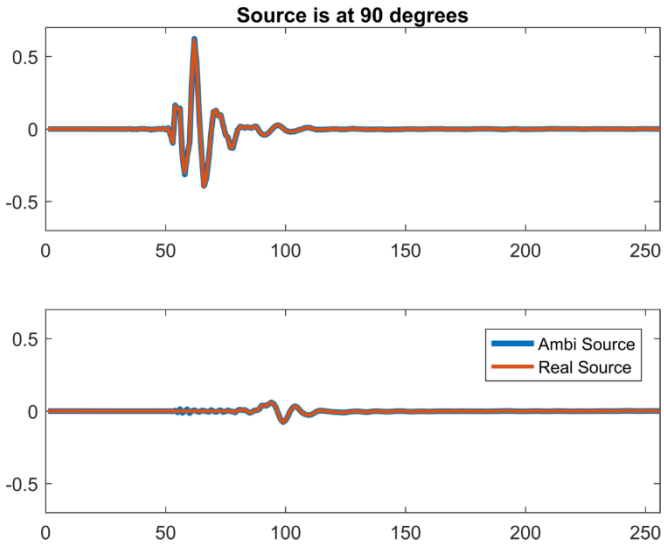

Figure 40: 17th Order HRIRs

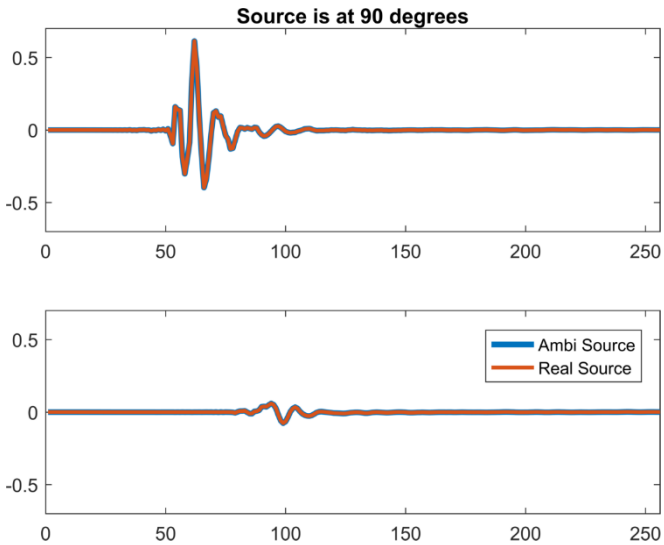

Figure 41: 35th Order HRIRs